August 2020

Exploratory Data Analysis and Point Process Modeling of Amateur Radio Spots

Nicole Chris
Clemson University, nicolechris2014@verizon.net

Exploratory Data Analysis and Point Process Modeling of Amateur Radio Spots

---

A Project
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Mathematical Sciences

---

by
Nicole Chris
August 2020

Accepted by:
Dr. Deborah Kunkel, Committee Chair
Dr. Xiaoqian Sun
Dr. Peter Kiessler

Abstract

Amateur radio spots are studied by scientists for many reasons. The Reverse Beacon Network (RBN) records thousands of spots and their characteristics on a daily basis. Located at the public server, http://www.reversebeacon.net/, it is open to be downloaded and explored by all. A 'spot' is by definition where a propagation path exists between a transmitter and a receptor location at a certain time and frequency (Miller et al., 2019). While this data can be useful to scientists, we do not have any knowledge to know when or how spots will occur. In this paper, we explore the idea of using the data for prediction. We start with the general question: Given input explanatory variables, what is the probability of a spot from a certain transmitter to a certain receptor? We begin with exploratory data analysis to find patterns or characteristics which may help with our choice of explanatory variables. Then, we research different statistical models and implement one which we deem most appropriate.

Table of Contents

Chapter 1

Introduction

As previously stated, the RBN records thousands of spots on a daily basis. The information recorded surrounding each spot includes: date and time (Coordinated Universal Time), frequency (kHz), bandwidth (m), decibel level, call signs, and the continents which the transmitter and receptor reside. The unique call signs were then mapped to a grid known as a maidenhead location. These maidenhead locations were obtain through private correspondence with collaborators. While some conceptual understanding of the physics behind amateur radio spots is left to the physicists, we will use our tools to help with exploratory data analysis and prediction. For this project, we work most directly with the following spot information: time, frequency, maidenhead locations, and continent locations.

Note that looking at each individual observation, it tells us where a spot occurred. Meaning at a certain time and frequency, we know that a spot between this specific transmitter and receptor happened. However, we do not have any instances where a spot did not happen. If a spot did not occur between a transmitter and receptor location at a certain time and frequency, does this mean the spot was impossible? In most cases, the answer is no. Therefore, we will use 'Presence-Only' to describe and model the RBN data. To do this we introduce the concept of a 'pseudo-absence' which by definition is a "random sample of sites taken from the population of interest where an absence is labeled to have occurred" (Ward et al., 2008). Then by definition 'Presence-Only' is a "combination of a sample of locations with known presences, and a background sample of locations from the whole population" (Ward et al., 2008). In the second half of this paper, we use these definitions to help us model the RBN data and find the best way to predict the probability of a spot occurring with

given information. Both parts of this project were completed using R Studio.

Chapter 2

Exploratory Data Analysis

This project is conducted using March 2019 - June 2019 data from the RBN ($\sim$ 443,000 observations). In addition to the given information, we append the following columns to the data set: latitude and longitude coordinates, distance, time zones, local time and season. The maidenhead locator system is a geographic coordinate system used to describe amateur radio operator locations. Each location corresponds to a two letter, 2 number, 2 letter sequence that allows scientists to determine its location. We use each maidenhead location to directly calculate the latitude and longitude of the transmitter and receptor, using ASCII values of the maidenhead characters in R. Since each maidenhead belongs to a very small grid on a world map, we achieve results that are within $\frac{1}{12}$ degree in longitude, and $\frac{1}{24}$ degree in latitude (Miller et al., 2019). The distance between the transmitter and the receptor is calculated using the Haversine formula, which given the latitude and longitude of two locations results in the shortest distance between the two points on a sphere. The equation is as follows:

$$d = 2r * \arcsin(\sqrt{\sin^2(\frac{\phi_1 - \phi_2}{2}) + cos(\phi_1)cos(\phi_2)\sin^2(\frac{\lambda_2 - \lambda_1}{2})}) \tag{2.1}$$

where d is the shortest distance, r is the radius of the earth, ϕ_1 and ϕ_2 are latitude of the two points, and λ_1 and λ_2 are the longitude of the two points. We use the *distHaversine* function in R for calculation. Time zone and local time were found through the latitude and longitude coordinates. Season was determined using the date of the spot. Local time is an important addition for the interpretation of any results based on time. Note that due to the timeline of the data, more

than 75% of our observations are from the spring season, and the rest are from the winter season. Therefore, we cannot come to any reasonable conclusion regarding this season variable. Additionally, for each unique transmitter location, we find their respective city, state, and country location. This helps create our visualization tool. Due to a much larger number of unique receptor locations, we do not follow the same procedure.

For general analysis, we find there to be 173 unique transmitter locations that come from ∼168 and ∼43 unique cities and countries respectively. There are 10,948 unique receptor locations. A continent analysis reveals that ∼86% of our unique transmitter locations are from North America and Europe. Additionally, ∼92% of our unique receptor locations are from the same two continents. South America, Africa, Asia, and Oceania have a much smaller number of unique locations. By definition, Oceania includes Australasia, Melanesia, Micronesia, and Polynesia.

The collaborators of (Miller et al., 2019) specify their interest in time of day and frequency as explanatory variables. They believe these two factors to have a large impact on the likelihood of a spot occurring. Additionally, the distance between the transmitter and receptor location could do the same. Therefore, we focusing on frequency, time and distance to look for patterns.

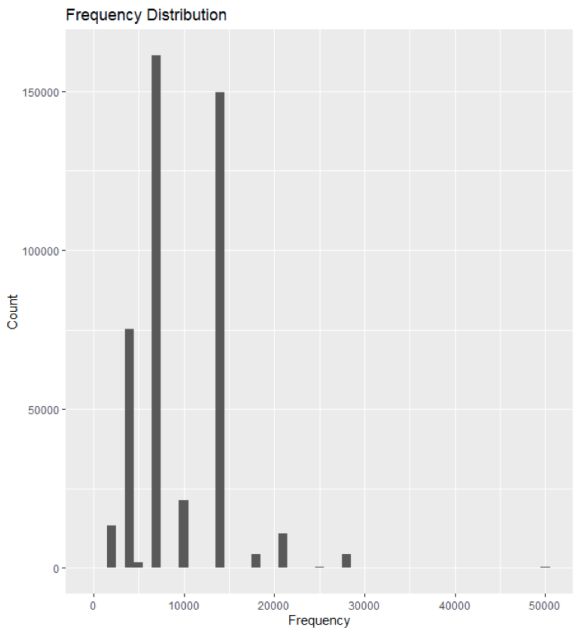

Figure 2.1: A histogram of frequency counts to reveal the most common frequencies and how these frequencies are distributed.

Figure 1 reveals that 7,000 kHz and 14,000 kHz are the most common frequencies. The third most common frequency is 3,500 kHz. Clearly these frequencies are not normally distributed, and only exist at certain levels. This is common as certain frequencies permit activities that others do not, so we should not expect a normally distributed histogram (Miller et al., 2019). This leads us to believe that a frequency specified at 7,000 kHz or 14,000 kHz may have a higher probability of occurring than a lower frequency or a very high frequency.

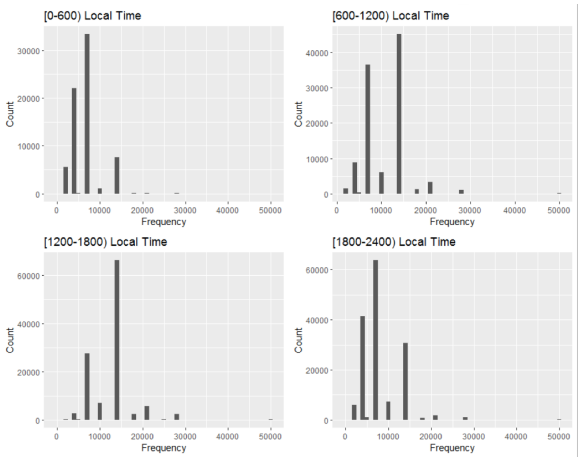

Figure 2.2: A histogram of frequency counts subsetted by four local transmitter times to reveal patterns based on frequency at different times of day. We classify [0-600) to be late night/early morning, [600-1200) as morning, [1200-1800) as afternoon/evening, and [1800-2400) to be night/late night.

Figure 2 reveals that lower frequencies, around 3000 kHz or below, have higher counts in the late night/early morning hours. As expected, 7,000 kHz are used consistently throughout the day, but are slightly less prevalent in the afternoon/evening hours. A similar pattern appears for 14,000 kHz, except that this frequency is slightly less prevalent in the late night/early morning hours. Much higher frequencies occur less often in the night/late night and late night/early morning hours. We may be inclined to believe that a spot at 50,000 kHz frequency and at 1:00 pm would have a higher probability of occuring than a spot at 50,000 kHz frequency and at 11:00 pm.

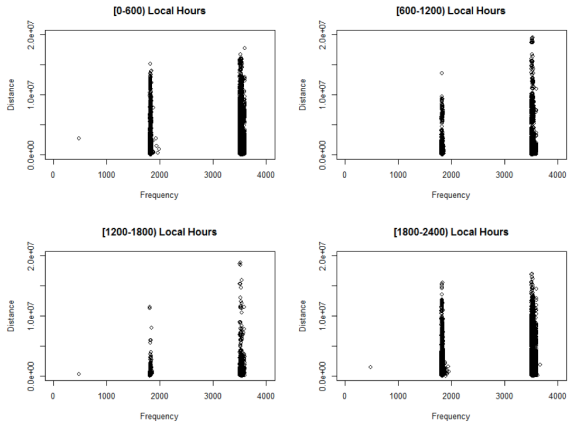

Figure 2.3: A scatterplot of frequencies (kHz) vs. distance (m) subsetted on the same four local transmitter times to reveal which frequencies travel shorter or farther distances and at what times of the day.

Figure 3 shows it is slightly more common for lower frequencies to travel further distances at night. This is most obviously seen in frequencies around 1,800 kHz. There is clearly more longer distanced calls in the late night/early morning hours than there are during the day. At 3,500 kHz we see a pattern of less distinction, but it still appears to follow a similar pattern of that of 1,800 kHz.

Chapter 3

Visualization Tool

We previously looked for patterns in variables regarding our entire data set. To build on this, a tool was created to subset a certain unique transmitter city and visualize what is going on at this specific location. This tool is built using tools from libraries *maps*, *ggplot2*, *tidyverse*, and *plyr*. We use this tool to look for any unique observations, and to examine whether the previously determined patterns follow for this location. We provide two analyzed examples.

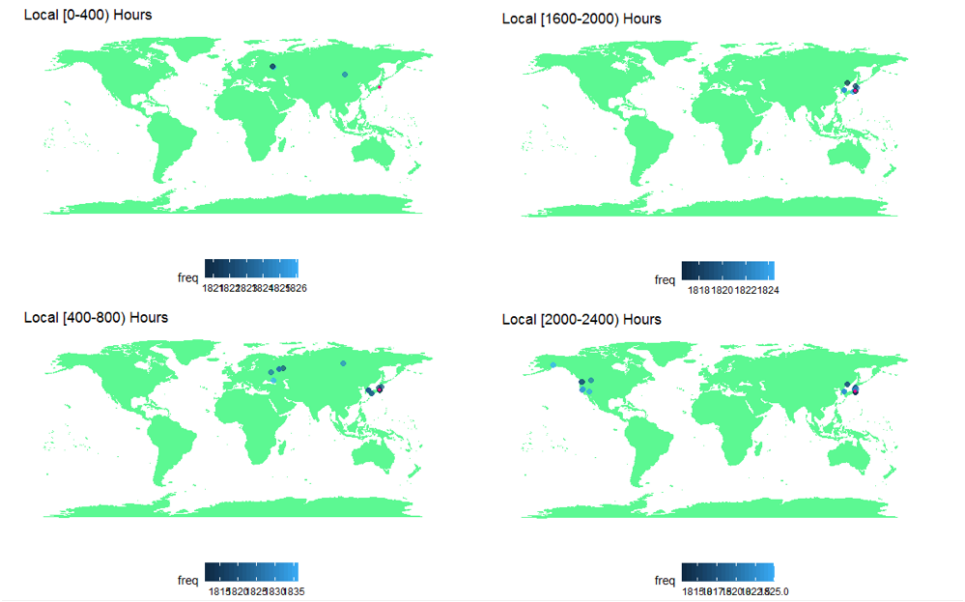

Figure 3.1: Visualization of transmitter in Suginami, Japan (pink), where receptor locations are color coded by frequency (blue).

.

From Figure 4 we see this transmitter location has few spots in relation to it. There are zero observations in [800-1200) and [1200-1800) hour groups. Observations during [1600-2000) hours seem to be strictly localized. Spots to Europe happen between [0-400) and [400-800) hours. Spots to the United States happen between [2000-2400). While we have no direct information regarding why spots occur between these places at these times, it is of interest to note. We see that all frequencies are below 2,000 kHz, and there is great distance on some of them. This location does support the pattern that lower frequencies travel further distances at night time.

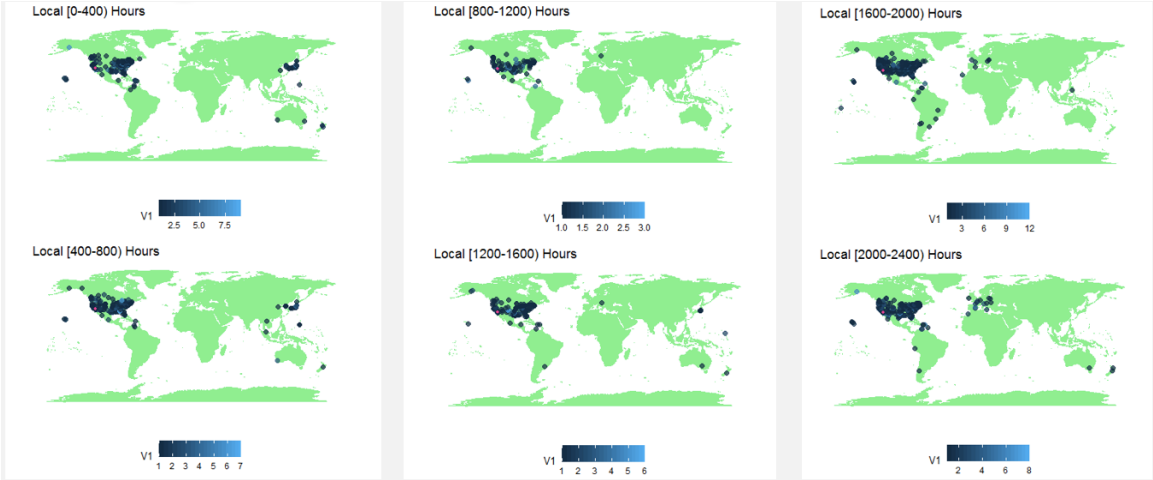

Figure 3.2: Visualization of transmitter in Rancho Cucamonga, California (pink), where receptor locations are color coded by their count (blue).

.

Figure 5 gives us different receptor information than Figure 4. Here we draw conclusions on receptor locations counts. Spots to Japan tend to happen in the [0-400) and [400-800) hour ranges. Each time frame has some locations with counts between 6-7, this leads me to believe there may be interesting information as to why these locations have significantly more spots than others.

This tool is not to be used to make any drastic conclusions in regards to explanatory variable patterns, but it can be used for more interests in specific transmitter location.

Chapter 4

Naive Method

Now that we have completed an initial explanatory analysis, we remind ourselves of the main goal of the project. Given certain characteristics, what is the probability of a spot occurring between a transmitter and receptor location? We know that we are classifying the RBN data as 'Presence-Only', so we begin to investigate models that will model this type of data and move us towards answering this general question. The first model we find is referred to as the Naive Model (Ward et al., 2008). We gain a solid understanding of this model and build from there. The process for the Naive Method is as follows:

- Take a population data set of size n, with covariates $\mathbf{X}$
- Calculate $\eta = \beta_0 + \beta_1 \mathbf{X}$
- Calculate $\mathbf{p} = \frac{e^{\eta}}{1+e^{\eta}}$
- Randomly generate presence (1) and absence (0) data in a vector $\mathbf{Y}$
- Apply a generalized linear model (glm) to find β_0 and β_1 estimates in $\mathit{logit}(\mathbf{Y}) = \beta_0 + \beta_1 \mathbf{X}$

Before applying this to our data, we run several simulations to examine the accuracy of the estimates in this process. After we apply the glm to the full presence and absence data set, we adjust the data set in multiple ways to compare their estimates to the true values. First, we eliminate all absence data. We predict the estimates to be well off in this scenario. After absence data is eliminated, we generate an m number of pseudo-absences and their covariates. We predict the slope estimates to converge to the true β_1 as m increases, while we expect the opposite of intercept

estimates (Ward et al., 2008). We repeat these adjustments, while also removing a selected number of presences from the data in addition to the removal of the absences. We ran several simulations, and provide one that encompasses the results that were very similar across all simulations. In this case, we randomly sampled a population size of n = 300 from the N(0,1) distribution, where $\beta_0 = 1$ and $\beta_1 = 2$. For each type of data, 20 simulations were run and the average estimates are recorded in the following table.

Data	m	$\hat{\beta}_0$	$\hat{\beta}_1$
Full	0	1.079	1.982
Removed Absences	0	26.57	$-1.19x10^{-8}$
Removed Absences	1,000	-1.681	.411
Removed Absences	10,000	-3.979	.378
Removed Absences, 50 presences	0	26.57	$-9.432x10^{-8}$
Removed Absences, 50 presences	1,000	-1.979	.434
Removed Absences, 50 presences	10,000	-4.278	.415

It is clear that the full data set has estimates closest to the true parameter values. It is also clear that both data sets with removed absences and zero added pseudo absences have results with very large errors. Finally, we note that our results do not converge closer to the true slope value as the number of pseudo-absences increases as suggested in (Ward et al., 2008). But we do conclude that the intercept is getting increasingly further from the true intercept value as the number of pseudo-absences increases as suggested in (Ward et al., 2008). Our last observation notes that our slope estimates are slightly closer to the true parameter value with 50 presences eliminated. With 1,000 pseudo-absences, the difference in slope estimates was .023. With 10,000 pseudo-absences, the difference in slope estimates was .037. Although we see a difference, the difference does not appear to be large enough to conclude that it is more helpful to remove any presences. We do not have enough evidence to conclude that all covariates are not important. Due to time constraints, we do not further explore why some of our results were not as we expected them to be, or whether we can make a definitive statement on whether all covariates are important. Given the limitations of the Naive Method, we do not directly use this methods on our radio spot data, therefore exploring these results will take place at a later time.

Chapter 5

EM Algorithm

Extended reading of (Ward et al., 2008) leads us to an advancement of the Naive Model known as an Expectation-Maximization Algorithm, also abbreviated EM Algorithm. This types of algorithm is used in many ways, to answer a wide variety of questions. In this paper, they use a version of the EM Algorithm for Presence-Only data. Its additions to the Naive Model are exactly laid out in the title, a maximization and an expectation step. Once we have our initial simulated presence and pseudo-absence data, $\mathbf{Y}$, the process begins with an initial population prevalence estimate, π. Where π is the $\frac{\text{Total number of presences in Y}}{\text{Total number of observations in Y}}$. From this step, the following are repeated until convergence:

1. *Maximization Step*:
   - Calculate $\hat{\eta}^{*(k)}$ for step k=1,2,... by fitting either
     a. logistic model of $\hat{\mathbf{y}}_{\mathbf{U}}^{k-1}$ on X
     b. logistic model of $\begin{pmatrix} \mathbf{1}_P \\ \mathbf{1}_U \\ \mathbf{0}_U \end{pmatrix}$ on $\begin{pmatrix} \mathbf{X}_P \\ \mathbf{X}_U \\ \mathbf{X}_U \end{pmatrix}$ with weights $\begin{pmatrix} \mathbf{1}_P \\ \hat{\mathbf{y}}_U^{k-1} \\ \mathbf{1}_U - \hat{\mathbf{y}}_U^{k-1} \end{pmatrix}$
   - Calculate $\hat{\eta}^{(k)} = \hat{\eta}^{*(k)} - log(\frac{n_p + \pi n_u}{(1-\pi)n_u}) - log(\frac{\pi}{(1-\pi)})$
2. *Expectation Step*:
   $\hat{y}_i^{(k)} = E[y_i|\hat{\eta}^k] = \frac{e^{\hat{\eta}^{(k)}}+1}{1+e^{\hat{\eta}^{(k)}}+1}$

Here, $\mathbf{U}$ is the combination of only the simulated presences and simulated absences. This

step information comes to us directly from (Ward et al., 2008). The added convergence step results in both the intercept and slope parameter estimates to be closer to the true values than that of the Naive Method (Ward et al., 2008). Although it has improved on the previous idea, there are still multiple limitations to this algorithm. The first is that we do not know the true population prevalence π. This is an initial estimate, and if it is not accurate could affect our results. Additionally, the proper chosen number of simulated pseudo-absences or presences, is not directly clear. In reference to our data specifically, the EM Algorithm is missing a spacial aspect that is very much important to the presence of radio spots. To combat these limitations, we decide not to simulate or apply this algorithm, and further papers are explored.

Chapter 6

Poisson Point Process Model

"Poisson Point Process Models Solve the Pseudo-Absence Problem for Presence-Only Data in Ecology" (Warton et al., 2010) directly confronts our issues with the EM algorithm. We spend time dissecting this paper in hopes to apply its ideas on our amateur radio spots. They propose analyzing Presence-Only data as a point process. This entails modeling the number of presence points n and their location y_i where $\mathbf{y} = (y_1, y_2, ..., y_n)$. For an inhomogeneous Poisson point process model, it is assumed that the locations of the presences are independent of one another. Then, the results are a function of intensity at a point, labeled λ_i. The interpretation is that λ_i is the expected number of presences per unit area (Warton et al., 2010). The resulting equation is:

$$log(\lambda_i) = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j \tag{6.1}$$

for covariate matrix $\mathbf{X}$ and parameter vector $\beta = (\beta_0, \beta_1, ..., \beta_m)$. For this model to work, the covariate values must be examined at all points in the window of area that the researcher is looking at. This is an important note for the application of this model.

In a similar way to pseudo-absences, this model chooses what they call quadrature points, which are associated with their own location denoted by $\mathbf{y}_0 = (y_{n+1}, ..., y_m)$. Each quadrature and presence point is assigned what is called a quadrature weight, w_i. Then $\mathbf{w} = (w_1, w_2, ..., w_n, w_{n+1}, ..., w_m)$. The number of quadrature points is chosen in convergence of the log-likelihood of the parameter

estimates:

$$l_{ppm}(\hat{\beta}; \mathbf{y}, \mathbf{y}_0, \mathbf{w}) = \sum_{i=1}^{m} w_i(z_i log(\lambda_i) - \lambda_i) \tag{6.2}$$

Where $z_i = \frac{I(i \in 1,\ldots,n)}{w_i}$. Quadrature weights can be determined in different ways, but are all calculated based on the area of the neighborhood A_i around each point y_i, such as the neighborhoods do not overlap, and the sum of the areas is the entire window where the data is explored (Warton et al., 2010).

A theorem proved by Warton et al. tells us that there is "Asymptotic equivalence of pseudo-absence logistic regression and Poisson point process models". Therefore, this solves our pseudo-absence concern from the EM algorithm. Additionally, we have a model that incorporates not only the presences, but also a spacial aspect surrounding presences and quadrature points. The interpretation of the Poisson point process method is much clearer than that of the previously researched methods. Due to these positive attributes, we adjust our original baseline question. Before we asked, based on given covariates, what is the probability of a spot occurring from a transmitter to a receptor? Now we ask, based on given covariates, how many receptors do we expect to receive a signal from a specific transmitter per unit area? This is a loaded question, and has some additional aspects to it, but it has an answer that makes more sense to us. Now that we have shifted the question, we feel comfortable applying the Poisson point process method to our amateur radio spot data to examine our results, and have a plan for the future of this project.

Chapter 7

Results

The following results came from using the R library *spatstat*, with its point process method function *ppm*. The data was imported as type *data.frame*, and was adjusted to be of type *point pattern*. We do this by using libraries *sf* and *sp* to sequentially turn our data into the following types, *data.frame*, *shapefile*, *SpatialPointsDataFrame*, finally to a *ppp* also known as a *point pattern*. To begin analysis, we subset all receptor locations into which continents they came from, and six local time sets. This gives us a total of 36 ppm's to run. To start, we do not subset on specific transmitters. That is, for now, we are only examining the expected number of locations per unit area to receive any signal at all (from any transmitter). For a pre-analysis, we examine the density plots of these ppm's to examine how the density of receptor locations changed over time. We include an example below.

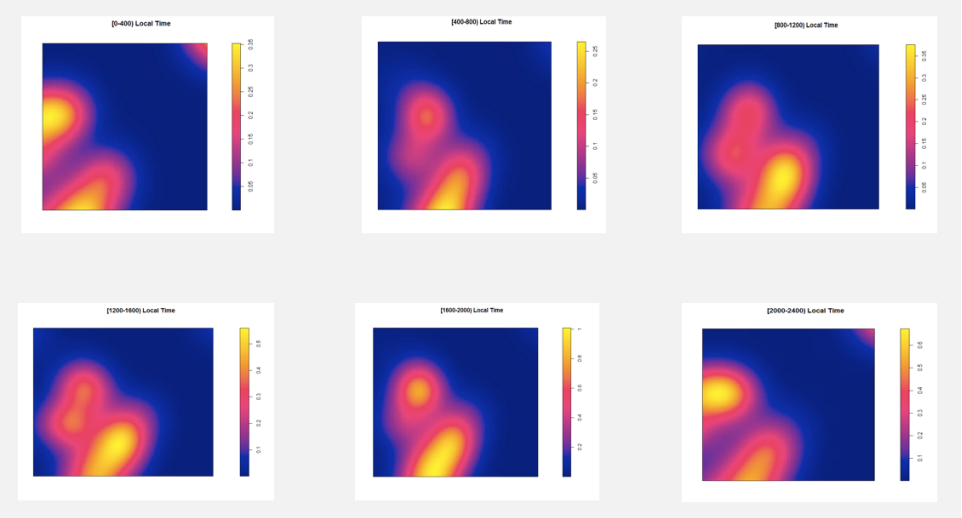

Figure 7.1: These plots include the densities of receptor locations in South America based on 6 subsetted local times.

.

Figure 6, which has an x-axis label of 'latitude' and a y axis of 'longitude', shows us how the density of receptor locations change in South America as time goes on throughout the day. The yellow spots are areas of higher density, pink spots are areas of middle density, and blue spots are areas of lower density. We see that throughout the day, the receptor locations in the northern part of the continent tend to become more spread out, and get more packed in the late night and early morning hours. Somewhat of the opposite appears to be true for the lower section of the continent. While this is a large area, it gives us an idea of what we are looking at, and how this model is going to work. If we subsetted an area in local time [0-400) where the density plot is mostly yellow, we might expect a higher intensity of receptor locations than that of the same local time, but subsetted on an area where the density plot is mostly blue.

For the model itself, we now run the ppm on each continent and local time. We begin with a stationary and homogeneous model. The results from these 36 ppm's are below.

Continent	[0-400)	[400-800)	[800-1200)	[1200-1600)	[1600-2000)	[2000-2400)
North America	.1533655	.4721778	.527795	.5063552	.9566693	1.267759
South America	.06477835	.04207356	.07155922	.1189929	.1972348	.1276183
Africa	.05369485	.03486471	.0576175	.0724377	.09215711	.1480479
Asia	.170256	.09577204	.1633424	.201462	.3801196	.2298187
Europe	1.13361	1.451765	1.713503	1.737206	2.605869	2.706535
Oceania	.01257624	.01480487	.01401256	.01783795	.04823126	.04309015

We see that "the expected number of receptors per unit area to receive a signal in North America from 12:00 am to 4:00 am is .1533655." This is a fairly low number, which makes sense in the fact that it is very early in the morning. Another interpretation is, "the expected number of receptors per unit area to receive a signal in Africa from 8:00 am to 12:00 pm is .0576175." This difference makes sense to us since we know that a much larger chunk of our data comes from North America and not as many observations come from Africa. We use this table to determine how the intensities change over time. This can aid us in future models. We see that 3 to 4 of the countries increase in somewhat of a monotone way throughout the day. However, some other have a pattern which appears to be higher in the afternoon or evening hours, and dips through the night time and morning hours. This could lead us to try introducing a quadratic fit into the model in regards to the time variable.

Chapter 8

Conclusions

Overall, our exploratory analysis showed that 7,000 kHz and 14,000 kHz are by far the most common frequencies, higher frequencies are more likely to produce spots during the day, and lower frequencies tend to travel further distances at night. Our Poisson point process model revealed that North America and Europe both have a higher expected number of receptor locations to be receiving a signal per unit area than all other continents throughout the day. About half of the continents experience a constant increase in intensity value from midnight through the rest of the day, while the other have experience an increased intensity value midday and a lower intensity value in the early morning and night time hours.

Chapter 9

Future Work

We start here with a decent basis. However, our stationary ppm leaves us with a lot of room to improve and new interesting questions. The immediate next step is to switch from subsetting based on time, to including time of spot as a covariate. From there, we would also like to add distance and frequency as covariates. This will give us more specific information regarding the intensities at different receptor locations, as opposed to this general time subsetted analysis. Then we need to incorporate much smaller areas. Continent conclusions can be interesting, and show us in the right direction, but the intensities per smaller area on a continent could vary widely. In this application, the results are said to be uniform across the continent, which we know would not be true. Therefore, subsetting on smaller areas would give us more specific and accurate results. The biggest issue to arise was the fact that covariates had to be known at all locations of the window. This is why we began with subsetting based on time, instead of including it as a covariate. This was an issue for frequency and distance as well. The goal, had there been more time to continue, was to create a tool that allows for subsetting based on one transmitter and one receptor area, and given time of day, frequency, and distance as covariates, determine the intensity for the receptor window. This goal will continue to be explored.

Bibliography

[1] Miller, E.S., et al. *Crowd-Sourced Observations of Day-to-Day Ionospheric Variability.* 2019.

[2] Warton, David I., and Leah Shepherd. "Poisson Point Process Models Solve the 'Pseudo-Absence Problem' for Presence-Only Data in Ecology." *The Annals of Applied Statistics*, vol.4, no.4, 2010, pp.2203-2204.

[3] Ward, Gill et al. "Presence-Only Data and the EM Algorithm." *Biometrics*, vol 65, no.2, 2008, pp.554-563

[4] Ferrier,Simon, et al. "Extended Statistical Approaches to Modelling Spatial Pattern in Biodiversity in Northeast New South Wales." *Biodiversity and Conservation*, 10 Aug. 2002, pp.2275-2307.

[5] Elith, Jane, et al. "A Statistical Explanation of MaxEnt for Ecologists." *Diversity and Distributions*, vol. 17, no. 1, 2010, pp. 43–57.